Trends in Bioinformatics

ISSN 1994-7941

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Trends in Bioinformatics

ISSN 1994-7941
DOI: 10.3923/tb.2018.17.24

Research Article
PremipreD: Precursor miRNA Prediction by Support Vector Machine Approach

[2]Sasti Gopal Das, [3]Hirak Jyoti Chakraborty and [1,2]Abhijit Datta

[1]DBT Centre for Bioinformatics, Presidency University, Kolkata, West Bengal, India
[2]Department of Botany, Jhargram Raj College, Jhargram, West Bengal, India
[3]Central Inland Fisheries Research Institute, Guwahati, India

Abstract

Background and Objective: Precursor microRNA expressions vary depending on their cellular environment and a large amount of genome segments can be folded in similar pseudo precursor's microRNA hairpins like structure. Therefore, detection of true precursor microRNA in a genome is challenging task. The computational prediction of precursor MicroRNAs first distinguishes a large amount of similar folded hairpins like structure in genome sequence as a pseudo or true precursor miRNAs. However, researchers need to be improving methods for identification of precursor MicroRNA in a genomic sequence. **Materials and Methods:** In this computational method, supervised machine learning approach was used as a classifier for classifying the true precursor miRNAs using sequence and secondary structure information. **Results:** The support vector machine (SVM) classifier achieved accuracy (Q) of 96.28% for predicting true pre-miRNAs. Here, a new precursor miRNA identification tool-PremipreD was developed which performs better in comparison to existing tools, in terms of overall performance and specificity. **Conclusion:** The PremipreD algorithm reduces the number of false positive prediction rate by using effective Support vector machine methods.

Key words: MicroRNA expression, gene expression, miRNA identification, support vector machine, PremipreD

Citation: Sasti Gopal Das, Hirak Jyoti Chakraborty and Abhijit Datta, 2018. PremipreD: Precursor miRNA prediction by support vector machine approach. Trends Bioinform., 11: 17-24.

Corresponding Author: Sasti Gopal Das, Department of Botany, Jhargram Raj College, 721507 Jhargram, West Bengal, India Tel: +91-9062300100

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

MicroRNAs (miRNAs) are ~22 nucleotide long with small non-coding single-stranded RNA molecules serving as master regulators of gene expression at the post-transcriptional level[1]. MicroRNAs are primarily transcribed by RNA polymerase II to produce primary miRNA with hairpin-like secondary structures[1]. The primary miRNAs are processed by the Drosha to generate precursor miRNA (Pre-miRNA) hairpin structures[1,2]. The hairpin pre-miRNAs are transported to the cytoplasm through the nuclear pore with the help of exportin-5 protein[3]. The pre-miRNA hairpin structures act as a structural motif for exportin-5 protein and also as a substrate for Dicer enzyme[1-3]. The precursor secondary structures of the miRNA are important for miRNA biogenesis. In the cytoplasm, pre-miRNAs are sliced by RNase III enzyme Dicer, which cuts them into ~22 nucleotide long miRNA duplexes[2]. The miRNA duplexes quickly convert into ~22 nucleotides long single-stranded small miRNAs. These miRNAs recognize their targets mainly by base-pairing interactions between the 5-end of the miRNA with the 3 -UTR of the genes[2]. The importance of miRNA research was rapidly increasing in molecular biology due to its impact on gene regulation. The MicroRNAs are associated with many gene regulation phenomena, such as those of neurodegenerative disorders, diabetes, cancer, cell development and cell death[4]. The MicroRNAs have low levels of expression and some are expressed in specific conditions only[4,5]. Due to these reasons, analysis of miRNA expressions with experimental studies is difficult. The miRNA sequences are found in non-coding sequences, making the identification of miRNA a technically challenging task in wet lab experiments[5]. The MicroRNA expression depends on various external factors such as cell type, physiological state of the organism and molecular tinkering of the gene[4,5]. Identification of miRNA discovery in next-generation sequencing (NGS) technology identifies thousands of gene expression in a single run, but miRNA expression depends on the biological condition[4,6]. The discovery of novel miRNAs from the organisms is still a great challenging task, therefore, a computational approach can help analyze the miRNA expression and miRNA detection[7].

Some researchers have developed computational methods for identifying miRNAs by using homology-based search and machine learning approaches. Homology-based approaches identify the conserved miRNAs from genomic sequences using BLAST search[8]. Many miRNA genes are not conserved, making the identification of non-conserved miRNAs from homology unfeasible.

Several authors suggested that pre-miRNA hairpin secondary structure and sequential information can assist in the computational identification of precursor miRNAs[9-11]. Machine learning approaches such as Support Vector Machine (SVM) can solve this problem. Some miRNA detection tools are developed based on machine learning algorithms, such as MiRenSVM[12], MiRPara[13], MiPred[14], ViralmiR[15], microPred[16], MiRFinder[17], iMcRNA-PseSSC[18], iMiRNA-PseDPC[19], miRNA-deKmer[20] and miRNA-dis[21], that looks for the miRNA gene. Most of the machine learning approaches uses known pre-miRNA as a positive data set and pseudo hairpins as a negative dataset for training their models. Some of the most important features of the machine learning techniques are nucleotide frequency, base-pairing interactions, structure properties and thermodynamic stability.

The vast number of sequences in the genome can fold into miRNA precursor like hairpin secondary structures[17]. The computational prediction of pre-miRNAs first distinguishes a genome sequence as a pseudo or true precursor miRNA[22]. However, researchers have been using limited techniques for identification of precursor miRNAs in genomic sequences. In this study, a machine learning approach was presented based on the support vector machine for identification of precursor miRNA genes in a given unknown precursor miRNAs secondary structures query sequence, called PremipreD, a reliable computational approach for improved identification of precursor miRNAs which can be used for identifying precursor miRNAs more efficiently.

MATERIALS AND METHODS

Research initial starting on March, 2016 and finalized by April, 2017.

Support vector machine dataset: The precursor miRNAs classifying system is trained with known precursor miRNAs as a positive dataset and the pseudo precursor miRNAs hairpins as a negative dataset. For positive miRNA precursors' datasets were downloaded from miRBase in 03/07/2014 (release 20) which contains 24521 precursor miRNA[23]. Duplicate pre-miRNAs are removed from the dataset and then randomly selected 10579 precursor miRNAs sequences from different species (animal, plant and virus). The Negative dataset is collected from Xue 8494 pseudo miRNAs[22], Zou 1446 pseudo miRNAs[24] and 639 tRNA from GtRNAdb 2.0[25,26]. The Negative dataset is selected by widely accepted pseudo-miRNAs characteristics: Minimum of 18 base pairings, maximum of -15 kcal mol^{-1} free energy and one loop[22]. To avoid imbalance problem, 10579 negative pseudo-miRNAs precursor sequences and 10579 positive miRNA precursors' sequences were selected.

Support vector machine feature extractions: Based on observations, the method focused on precursor miRNAs secondary structural conformation and sequence information[12,22,24]. In present study, we used three featured category for classifying the pre-miRNAs and pseudo pre-miRNAs. In the first featured category, we focused on secondary structure conformation based on class intervals[12]. In the second featured category triplet structure-sequence elements information were considered [22]. The third featured category focused on Tri-nucleotides sequence composition information[23]. The miRNA precursor secondary structures are predicted using RNAfold (Vienna package)[27]. Precursor miRNA secondary structures are represented in dot-bracket notation.

Secondary structural conformation feature: Precursor miRNA secondary structural pattern such as the intramolecular base pairing of precursor miRNA is a significant important feature for classifying the precursor miRNAs[12]. The discriminative powers of the two different class of precursor miRNAs secondary structural conformation (dot-bracket notation) are analyzed[27]. Precursor miRNA secondary structural feature such as Minimum Free Energy, Watson-crick base pairing (AU, GC), Wobble base pairing (G-U) and unpaired bases (A, G, C, U) is divided by sequence length. The precursor miRNA feature is analyzed in the class interval based in F-score method as can be seen in Eq. 1[12]. The discriminatory power of each precursor miRNA secondary structural conformation feature F-score was shown in Table 1 and the 13 most important SVM features F-score are selected and the selected SVM features are highlighted in bold.

Triplet structure-sequence elements features: The RNAFold predicted dot-bracket notation secondary structure, paired as "(" and unpaired as ".", respectively. In triplet structure-sequence feature selections for precursor miRNA, three contiguous structures information "(((", "((.", "(..", "(.(", ".((", ".(.", "..(" ,"..." and one middle nucleotides among the 3 information in bold "xAx", "xUx", "xGx", etc "xCx" , there are 32 (8×4) possible structure-sequence combinations, which we denote as "U(((", "A((." and so on are shown in (Table 2). For each 32 triplet structure-sequence, features were analysed using F score (Eq. 1). The 11 most important SVM feature are selected and selected SVM features were highlighted in bold[22].

Tri-nucleotide sequence composition features: In the third method, we focus on the Tri-nucleotide sequence composition of precursor miRNA. For Tri-nucleotide feature selection, we analyzed the discriminative powers of the two different class of precursor miRNA sequence[24]. 64 Tri-nucleotide composition elements are analyzed using F-score described in Eq. 1. Table 3 showed that the most discriminative 10 tri-nucleotide SVM features are selected and the selected SVM features are highlighted in bold.

Feature selection: The F-score is a feature selection technique which measures the discriminate power of two sets of real numbers. In general, most important feature selections of two real datasets information of features lead to production of a better performance of model. In training vectors x_k, k = 1,... m, if the number of true positive and true negative instances are n+ and n^- , respectively, then the F-score of the i-th feature was defined in Eq. 1:

$$f(i) = \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n+}\left(\bar{x}_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n-}\left(\bar{x}_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2} \quad (1)$$

where, $\bar{x}_i$ is a whole, $\bar{x}_i^{(+)}$ is a true positive and $\bar{x}_i^{(-)}$ is a true negative are the average of the i-th feature of the data sets, respectively, $\bar{x}_{k,i}^{(+)}$ is the i-th feature of the k-th true positive instance and $\bar{x}_{k,i}^{(-)}$ is the i-th feature of the k-th true negative instance[16]. The F-score threshold was chosen based on SVM model accuracy using various cut-off F-score. Using a cut-off F-score, the retained features are 102, 92, 73, 66, 34 and 31 and the corresponding accuracy of the SVM model are 91.94, 92.39, 93.65, 94.13, 96.28 and 95.41%, respectively. F-score greater than 34 features give better accuracy. Therefore it was decided to use 34 precursor miRNA features as the best feature for our SVM model.

Table 1: Frequency of secondary structural conformation feature in class interval (category 1)

Feature name	1-10	11-20	21-30	31-40	41-50	51-60	61-100
Pre-miRNA free energy	0.0000	0.0094	0.0686	0.0013	0.0405	0.0180	0.0047
Total base pairing (AU,GC,GU)	0.0000	0.0001	0.0903	0.0146	0.1686	0.0002	0.0000
Unpaired bases adenine (A)	0.0946	0.0629	0.0080	0.0000	0.0000	0.0000	0.0000
Unpaired bases guanine (G)	0.0536	0.0477	0.0008	0.0000	0.0000	0.0000	0.0000
Unpaired bases uracil (U)	0.0052	0.0027	0.0013	0.0000	0.0000	0.0000	0.0000
Unpaired bases cytosine (C)	0.1142	0.0915	0.0043	0.0000	0.0000	0.0000	0.0000
Base pairing (AU)	0.3680	0.0985	0.0669	0.0002	0.0000	0.0000	0.0000
Base pairing (GC)	0.0031	0.0036	0.0152	0.0000	0.0000	0.0000	0.0000
Wobble base pairing (G-U)	0.0000	0.0016	0.0000	0.0000	0.0000	0.0000	0.0000

Table 2: Frequency of triplet structure-sequence elements features (category 2)

A(((= 0.4424	U(((= 0.4664	G(((= 0.1192	C(((= 0.0661
A((. = 0.0923	U((. = 0.0853	G((. = 0.0408	C((. = 0.0431
A(.(= 0.0157	U(.(= 0.0590	G(.(= 0.0015	C(.(= 0.0376
A(.. = 0.0264	U(.. = 0.0058	G(.. = 0.0300	C(.. = 0.0402
A.((= 0.0728	U.((= 0.0661	G.((= 0.0457	C.((= 0.0145
A.(. = 0.0000	U.(. = 0.0003	G.(. = 0.0193	C.(. = 0.0113
A..(= 0.0289	U..(= 0.0038	G..(= 0.0317	C..(= 0.0450
A... = 0.0148	U... = 0.0000	G... = 0.1380	C... = 0.1512

Table 3: Frequency of tri-nucleotide sequence composition features (category 3)

AAA = 0.0861	UAA = 0.1305	GAA = 0.0292	CAA = 0.0547
AAU = 0.1520	UAU = 0.1924	GAU = 0.0698	CAU = 0.0696
AAG = 0.0241	UAG = 0.0779	GAG = 0.0011	CAG = 0.0397
AAC = 0.0397	UAC = 0.0753	GAC = 0.0103	CAC = 0.0007
AUA = 0.1571	UUA = 0.1606	GUA = 0.0747	CUA = 0.0959
AUU = 0.2030	UUU = 0.1865	GUU = 0.1362	CUU = 0.0572
AUG = 0.0798	UUG = 0.1827	GUG = 0.0162	CUG = 0.0126
AUC = 0.0470	UUC = 0.0584	GUC = 0.0037	CUC = 0.0023
AGA = 0.0116	UGA = 0.0893	GGA = 0.0059	CGA = 0.0051
AGU = 0.0663	UGU = 0.1627	GGU = 0.0003	CGU = 0.0000
AGG = 0.0192	UGG = 0.0025	GGG = 0.0381	CGG = 0.0318
AGC = 0.0023	UGC = 0.0121	GGC = 0.0597	CGC = 0.0199
ACA = 0.0345	UCA = 0.0497	GCA = 0.0019	CCA = 0.0240
ACU = 0.0515	UCU = 0.0476	GCU = 0.0000	CCU = 0.0172
ACG = 0.0006	UCG = 0.0000	GCG = 0.0154	CCG = 0.0556
ACC = 0.0113	UCC = 0.0190	GCC = 0.0605	CCC = 0.0789

Support vector machines and SVM model selection: For the proper training of SVM models, selection of several hyper parameters is an essential prerequisite, their parameters values determine the function that SVM optimizes and it has a significant impact on the performance of the trained SVM classifiers. The best hyper-parameter set was selected by a 10-fold cross-validation on the dataset. The widely used radial basis function (RBF) kernel was chosen for training of the model[12]. The SVM learns from the transformation of the input data into another higher dimensional feature space for accurate classification. The training input vector corresponds to the precursor miRNA specific features such as {+1,-1} (+1 for true precursor miRNAs, -1 for pseudo precursor miRNAs). We built SVM model based on the 34 features. Therefore, all the positive and negative precursor miRNAs from the training set are implicitly mapped from the input space to a feature space determined by the RBF kernel. In this feature space an optimal hyper-plane is learned by the SVM. In this regard, a suitable setting of the SVM parameter *C* and the RBF kernel parameter gamma (g) are determined with a 10-fold cross validation on the training dataset. The parameter C is used to control the trade-off between the training errors and parameter gamma controls the width of the RBF kernel. From the experiment, the best set of parameters for 34 feature set were obtained as C = 10 and g = 0.2[12].

These performance metrics are defined as follows:

- Sensitivity (SE) = TP/TP+FN
- Specificity (SP) = TN/TN+FP
- Accuracy (Q) = TP+TN/TP+TN+FP+FN

RESULTS AND DISCUSSION

Performance measures for the SVM model: The performance of this method was measured by the total number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), sensitivity (SE), specificity (SP) and accuracy (Q) and area under the curve (AUC)[12-17]. In order to assess the performance of models developed in this study, various parameters like sensitivity, specificity and accuracy (Q) and area under the curve (AUC) were calculated as shown in Table 4. Performance measurement of SVM model tested with data from miRBase in 03/07/2014 (release 20)[23] in ten-fold cross validation. About 90% of the dataset was employed for training the SVM models and the remaining independent 10% was used for testing. This process was iterated 10 times. The predictive performance of classifiers was evaluated by an SVM threshold zero.

Comparison of PremipreD prediction abilities of precursor miRNAs using completely independent test data with existing prediction tools: The predictive ability of this method (PremipreD) was compared to four other current miRNA predictions software packages iMcRNA-PseSSC[18], iMiRNA-PseDPC[19], miRNA-deKmer[20] and miRNA-dis[21]. The software was tested with randomly selected 500 positive, completely independent test data set, which were retrieved from new releases miRBase (v 21)[23]. About 500 pre-miRNA-like hairpins negative test data set was built from (SinEx DB: A database for single exon coding sequences in mammalian genomes)[28]. This database curetted Eukaryotic 'single exon genes' protein-coding sequence (CDS). These protein-coding sequence (CDS) segments, though they are similar to pre-miRNAs structures but have not been reported as pre-miRNAs. For creating negative test data, widely accepted characteristics were used: minimum of 18 base pairings, maximum of -15 kcal/mol free energy and one loops are taken[22]. Ability to correctly predict experimentally verified known precursor miRNA and pre-miRNA-like hairpins negative test data set is tested, PremipreD outperformed all other packages (Table 5).

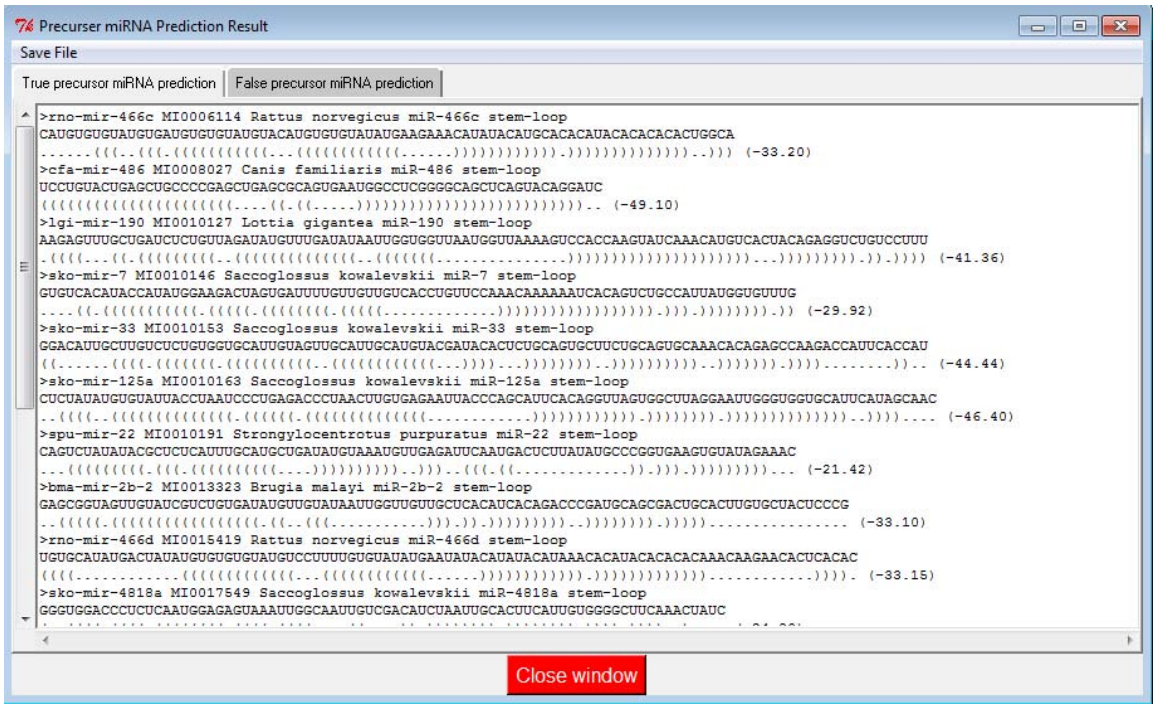

Fig. 1: Screenshot of PremipreD tool GUI give precursor miRNA output

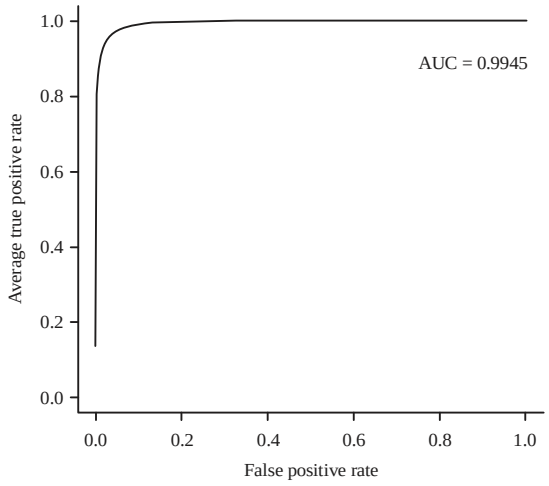

Fig. 2: ROC curve denotes the performance of this method, the Y-axis represents the true positive rate (sensitivity) and X-axis shows the false positive rate (1-specificity)

Table 4: Performance of the SVM model in ten-fold cross validation

TP	FN	TN	FP	Sensitivity	Specificity	Accuracy	AUC
10294	285	10077	502	0.9731	0.9525	0.9628	0.9945

A large amount of miRNAs could not be identified in many species. However, PremipreD can distinguish the true pre-miRNAs (Fig. 1). In the modern research, precursor miRNA identification is the most important part for molecular biology due to its impact on gene regulation. Existing prediction method can predict thousand of miRNA genes by homology search, but researchers need better and more efficient methods for predicting true precursor miRNA. In genomic sequences, many sequence segments can fold into the hairpin-like secondary structure, so distinguishes segment into true miRNA or Pseudo precursor miRNA gene. The result showed that the SVM model can predict 96.28% data correctly and ROC curve 0.9945 (Fig. 2) indicate that this algorithm accurately predicts pre-miRNA. In this study, we present PremipreD (Fig. 3), a computational method that predicts true precursor miRNA by utilizing SVM model on the basis of the secondary structure of miRNA (Fig. 4). PremipreD method is simpler than the existing methods, irrespective of any specific class of precursor miRNA. The same feature filtering and SVM parameter optimization steps were performed. The SVM feature model parameter was used for training the SVM classifier.

Table 5: Comparison of prediction abilities of PremiPreD with other existing tools

Precursor miRNA prediction software	TP	FN	TN	FP	Sensitivity	Specificity	Accuracy
iMcRNA-PseSSC	448	52	416	84	0.896	0.832	0.864
miRNA-deKmer	366	134	371	129	0.732	0.742	0.737
miRNA-dis	479	21	402	98	0.958	0.804	0.881
iMiRNA-PseDPC	477	23	414	86	0.954	0.828	0.891
PremipreD	445	55	461	39	0.890	0.922	0.906

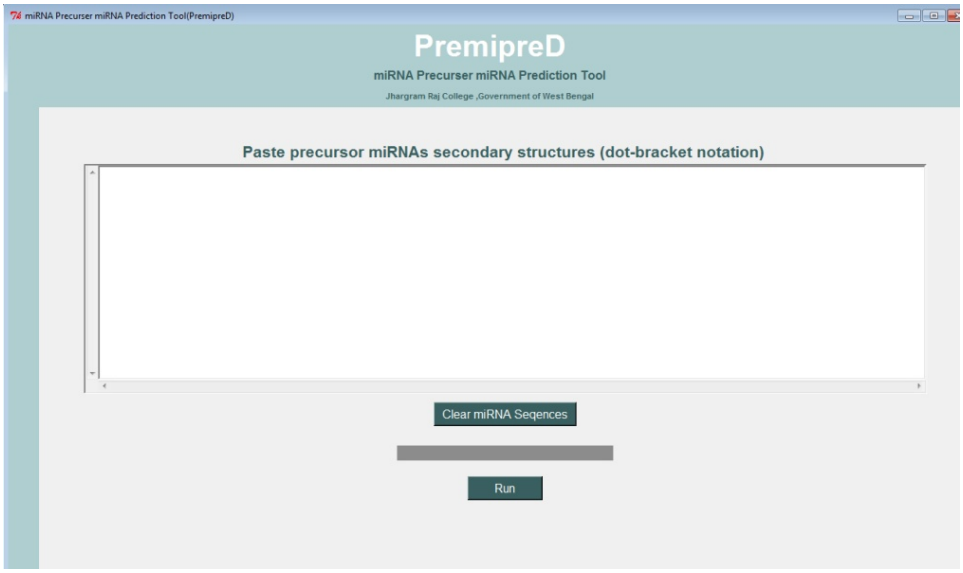

Fig. 3: Screenshot of PremipreD tool GUI Input. User submits unknown precursor miRNAs secondary structures in the input box

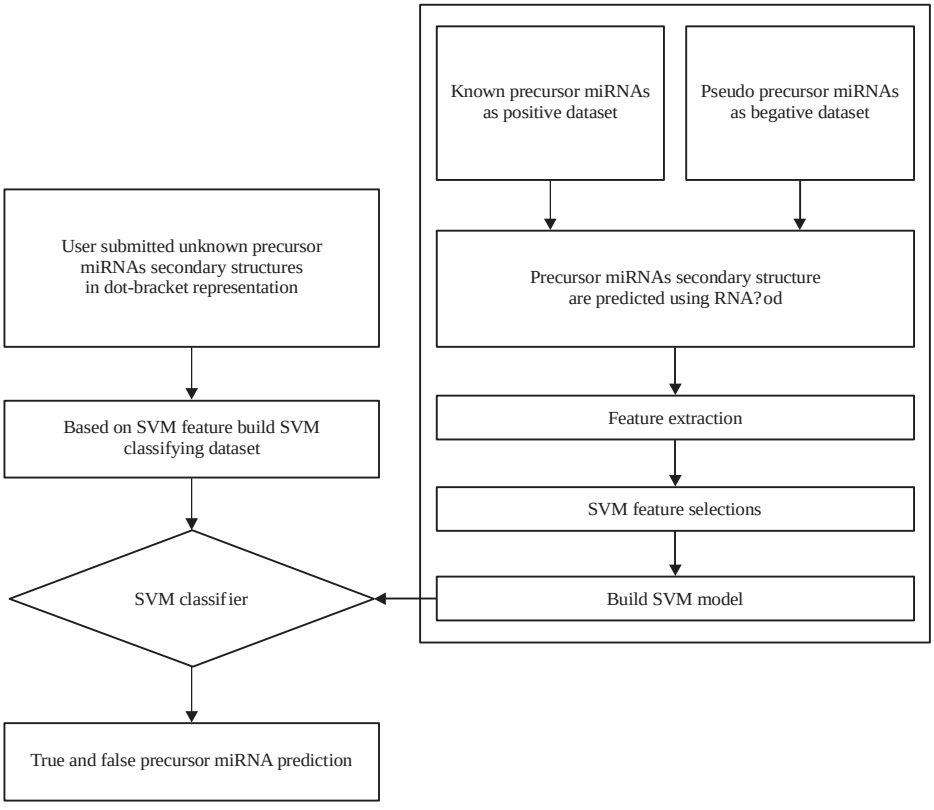

Fig. 4: Workflow of the precursor miRNA prediction method in PremiPred

CONCLUSION

The PremipreD classifier can predict novel precursor miRNA from animal, plant and viruses. The comparison of prediction abilities between PremipreD and the existing methods, in terms of specificity and overall accuracy indicate that the overall performance of the PremipreD algorithm significantly outperforms all other tools.

SIGNIFICANCE STATEMENTS

This study predicts unknown precursor miRNA from genomic sequences that can be beneficial for finding the true precursor miRNAs in genomic sequences. This study will help the researcher find out the uncover precursor miRNA that many researchers were not able to explore. Thus a new method on precursor miRNA predicted model could be supportive to understand the characteristics precursor miRNA involved in miRNA biogenesis.

ACKNOWLEDGMENT

The authors remain grateful to the DBT, BIF-BTBI scheme, GoI for providing the infrastructure. The authors would like to thank Aditi Gangopadhyay (DBT Centre for Bioinformatics, Presidency University, Kolkata) for her help during manuscript preparation.

REFERENCES

1. Kim, V.N., J. Han and M.C. Siomi, 2009. Biogenesis of small RNAs in animals. Nature Rev. Mol. Cell Biol., 10: 126-139.
2. Ruby, J.G., A. Stark, W.K. Johnston, M. Kellis, D.P. Bartel and E.C. Lai, 2007. Evolution, biogenesis, expression and target predictions of a substantially expanded set of *Drosophila* microRNAs. Genome Res., 17: 1850-1864.
3. Gulyaeva, L.F. and N.E. Kushlinskiy, 2016. Regulatory mechanisms of microRNA expression. J. Transl. Med., Vol. 14. 10.1186/s12967-016-0893-x.
4. Ryu, S., N. Joshi, K. McDonnell, J. Woo and H. Choi *et al*., 2011. Discovery of novel human breast cancer microRNAs from deep sequencing data by analysis of pri-microRNA secondary structures. Plos One, Vol. 6. 10.1371/journal.pone.0016403.
5. Melamed, Z.E., A. Levy, R. Ashwal-Fluss, G. Lev-Maor and K. Mekahel *et al*., 2013. Alternative splicing regulates biogenesis of miRNAs located across exon-intron junctions. Mol. Cell, 50: 869-881.
6. Friedlander, M.R., W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel and N. Rajewsky, 2008. Discovering microRNAs from deep sequencing data using miRDeep. Nature Biotechnol., 26: 407-415.
7. Stark, A., M.F. Lin, P. Kheradpour, J.S. Pedersen and L. Parts *et al*., 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. Nature, 450: 219-232.
8. Gupta, H., T. Tiwari, M. Patel, A. Mehta and A. Ghosh, 2015. An approach to identify the novel miRNA encoded from *H. Annuus* EST sequences. Genomics Data, 6: 139-144.
9. Wan, Y., M. Kertesz, R.C. Spitale, E. Segal and H.Y. Chang, 2011. Understanding the transcriptome through RNA structure. Nature Rev. Genet., 12: 641-655.
10. Auyeung, V.C., I. Ulitsky, S.E. McGeary and D.P. Bartel, 2013. Beyond secondary structure: Primary-sequence determinants license Pri-miRNA hairpins for processing. Cell, 152: 844-858.
11. Jinek, M. and J.A. Doudna, 2009. A three-dimensional view of the molecular machinery of RNA interference. Nature, 457: 405-412.
12. Ding, J., S. Zhou and J. Guan, 2010. MiRenSVM: Towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. BMC Bioinfor., Vol. 11. 10.1186/1471-2105-11-S11-S11.
13. Wu, Y., B. Wei, H. Liu, T. Li and S. Rayner, 2011. MiRPara: A SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. BMC Bioinfor., Vol. 12. 10.1186/1471-2105-12-107
14. Jiang, P., H. Wu, W. Wang, W. Ma, X. Sun and Z. Lu, 2007. MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res., 35: W339-W344.
15. Huang, K.Y., T.Y. Lee, Y.C. Teng and T.H. Chang, 2015. ViralmiR: A support-vector-machine-based method for predicting viral microRNA precursors. BMC Bioinfor., Vol. 16. 10.1186/1471-2105-16-S1-S9.
16. Batuwita, R. and V. Palade, 2009. MicroPred: Effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics, 25: 989-995.
17. Huang, T.H., B. Fan, M.F. Rothschild, Z.L. Hu, K. Li and S.H. Zhao, 2007. MiRFinder: An improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinfor., Vol. 8. 10.1186/1471-2105-8-341.
18. Liu, B., L. Fang, F. Liu, X. Wang, J. Chen and K.C. Chou, 2015. Identification of real microRNA precursors with a pseudo structure status composition approach. Plos One, Vol. 10. 10.1371/journal.pone.0121501.
19. Liu, B., L. Fang, F. Liu, X. Wang and K.C. Chou, 2016. iMiRNA-PseDPC: MicroRNA precursor identification with a pseudo distance-pair composition approach. J. Biomol. Struct. Dynamics, 34: 223-235.
20. Liu, B., L. Fang, S. Wang, X. Wang, H. Li and K.C. Chou, 2015. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. J. Theor. Biol., 385: 153-159.
21. Liu, B., L. Fang, J. Chen, F. Liu and X. Wang, 2015. MiRNA-dis: MicroRNA precursor identification based on distance structure status pairs. Mol. BioSyst., 11: 1194-1204.
22. Xue, C., F. Li, T. He, G.P. Liu, Y. Li and X. Zhang, 2005. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinfor., Vol. 6. 10.1186/1471-2105-6-310.

23. Kozomara, A. and S. Griffiths-Jones, 2011. miRBase: Integrating microRNA annotation and deep-sequencing data. Nucleic. Acids Res., 39: D152-D157.
24. Wei, L., M. Liao, Y. Gao, R. Ji, Z. He and Q. Zou, 2014. Improved and promising identification of human microRNAs by incorporating a high-quality negative set. IEEE/ACM Trans. Comput. Biol. Bioinfor., 11: 192-201.
25. Chan, P.P. and T.M. Lowe, 2008. GtRNAdb: A database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res., 37: D93-D97.
26. Tempel, S., B. Zerath, F. Zehraoui and F. Tahi, 2015. MiRBoost: Boosting support vector machines for microRNA precursor classification. RNA, 21: 775-785.
27. Lorenz, R., S.H. Bernhart, C.H. Zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler and I.L. Hofacker, 2016. ViennaRNA package 2.0. Algorithms Mol. Biol., Vol. 6. 10.1186/1748-7188-6-26.
28. Jorquera, R., R. Ortiz, F. Ossandon, J.P. Cardenas, R. Sepulveda, C. Gonzalez and D.S. Holmes, 2016. SinEx DB: A database for single exon coding sequences in mammalian genomes. Database, Vol. 2016. 10.1093/database/baw095/2630477.